Center Wide Bandwidth Calculations

Rob Farber

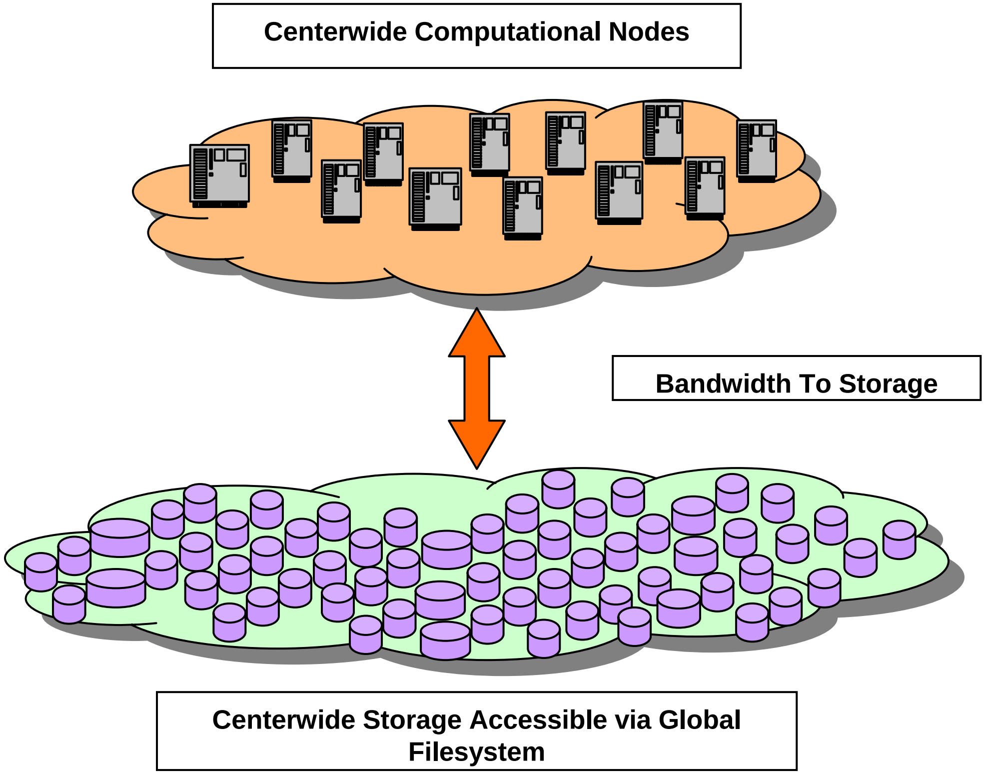
Special Thanks to Craig Tull and Nancy Meyer

Outline

- Present two background slides
- Talk about the goals of the model
- Introduce the streaming model
- Express the model as an Excel spreadsheet
- Apply it (as time permits) to a:
  - Simple example
  - Current example
  - Future example
  - Cost analysis
- Summarize/Future Directions
- Questions?

Determining Streaming Bandwidth Has Been A GUPFS Focus

- Two streaming IO benchmarks are commonly used
  - Pioraw
  - Mptio
- *Look for Red Flags as streaming performance requires sufficient Meta-data performance, data coherency and low-latency IO, as measured by:*
  - Metabench
  - smallFileTest



GUPFS Has Been Collecting Test Results For Nearly Two Years

- Lots of data
  - In many different forms
  - Located in many different places
- Numerous meetings have occurred
  - Within LBL about how groups are:
    - Forecasting their future needs.
    - Planning procurement to meet those needs .
  - Externally with vendors
    - About current and future technologies
    - To gain insight about where future “sweet spots” will be for component price and performance.

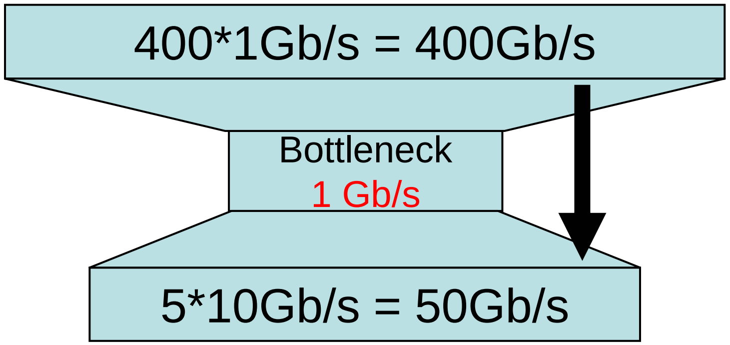
HOW TO INTEGRATE AND COMMUNICATE ALL THIS INFORMATION?

Goals Of The Centerwide Bandwidth Model

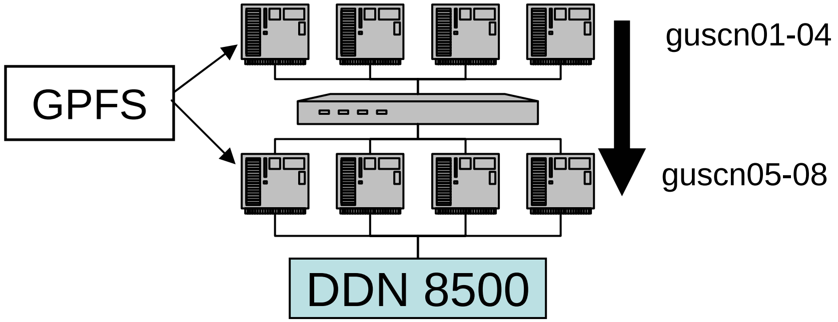
- **Integrate Information** (to facilitate decision makers)
  - Of existing measurements and projections into GUPFS "What-if" scenario analysis, which can provide guidance on:
    - Component selection
      - Type
      - Number
    - Bottleneck identification
    - Cost estimation (as required)
- **Communicate Scenarios**
  - For Funding and Procurement
    - Justify component selection and performance numbers.
    - Provide cost estimates
    - Define a roadmap to meet anticipated needs
  - In a way people can understand
    - Utilize Excel spreadsheets as management and funding agencies are familiar with them.
    - Use a simple model: "just enough and no more".

Use A Simple Model

- Streaming IO is similar to a plumbing problem
  - Very large pipes carry more than small pipes (effective data rate is the dominant effect).
  - Lots of small pipes transfer more than a few small pipes (data rates are additive).
  - A minimum transfer rate at any point in the flow, or bottleneck, limits the maximum throughput.
- The GUPFS streaming benchmarks and other measurements determine the effective data rate (i.e. size of the pipe) for various components and file-systems.



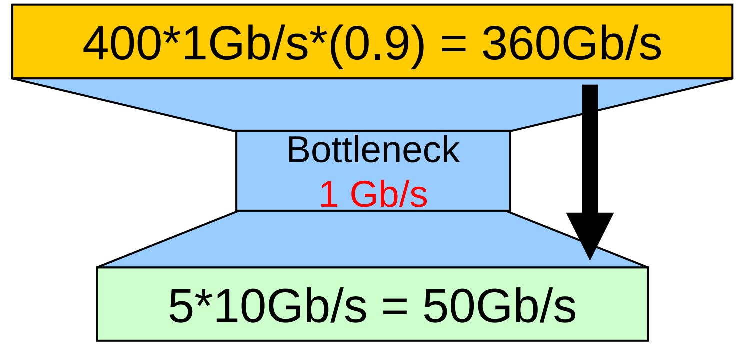
Streaming throughput of this schematic “system” is 1 Gb/s



“pioraw” streaming throughput of this system is 100-227 MB/s (1-4 clients)

Express The Model As A Spreadsheet

- Conventions:
  - Bottlenecks are highlighted in **Red**.
  - **Blue** brings attn to user changes.
  - Computed cells are colored and bold.
- Use natural units to express device characteristics (Gb/s, number of units, …).
  - Add device detail (i.e. transport efficiency) only as necessary.
- Automatically identify bottlenecks.



Streaming throughput of this schematic “system” is 1 Gb/s

Year	Number of Compute Nodes	Node Link Speed (Gb/s)	Number of Ports	Efficiency (%) of transport	Effective BW per Node (Gb/s)	Aggregate BW from Compute Nodes (Gb/s)	Bottleneck (Gb/s)	Aggregate BW to Storage (Gb/s)	Num Storage Units	Storage BW (Gb/s)
2004	400	1	1	90%	**0.9**	**360**	1	**50**	5	10
2005	400	1	1	90%	**0.9**	**360**	1000	**50**	5	10
2006	400	1	1	90%	**0.9**	**360**	1000	**500**	5	100

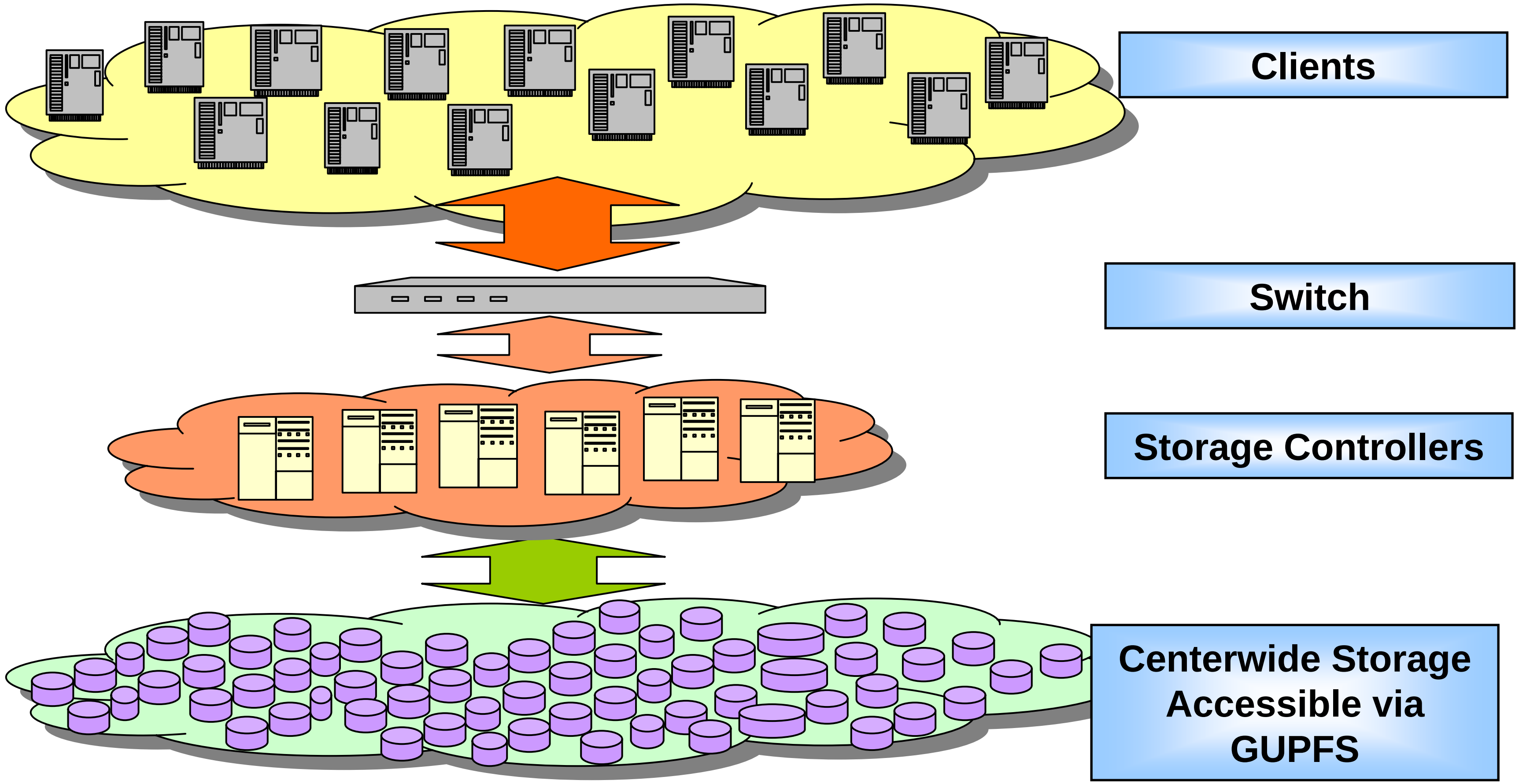
Apply It: Let's Talk About Some Scenarios Focusing On Our Goals

Scenarios:

1. Simple:
   - Briefly discuss a single cluster evolving over time.
2. Current:
   - Seaborg
   - PDSF
   - HPSS
3. Possible Future Center:
   - Seaborg
   - PDSF
   - HPSS
   - Sys1
   - Sys2
   - Sys3
   - Sys4

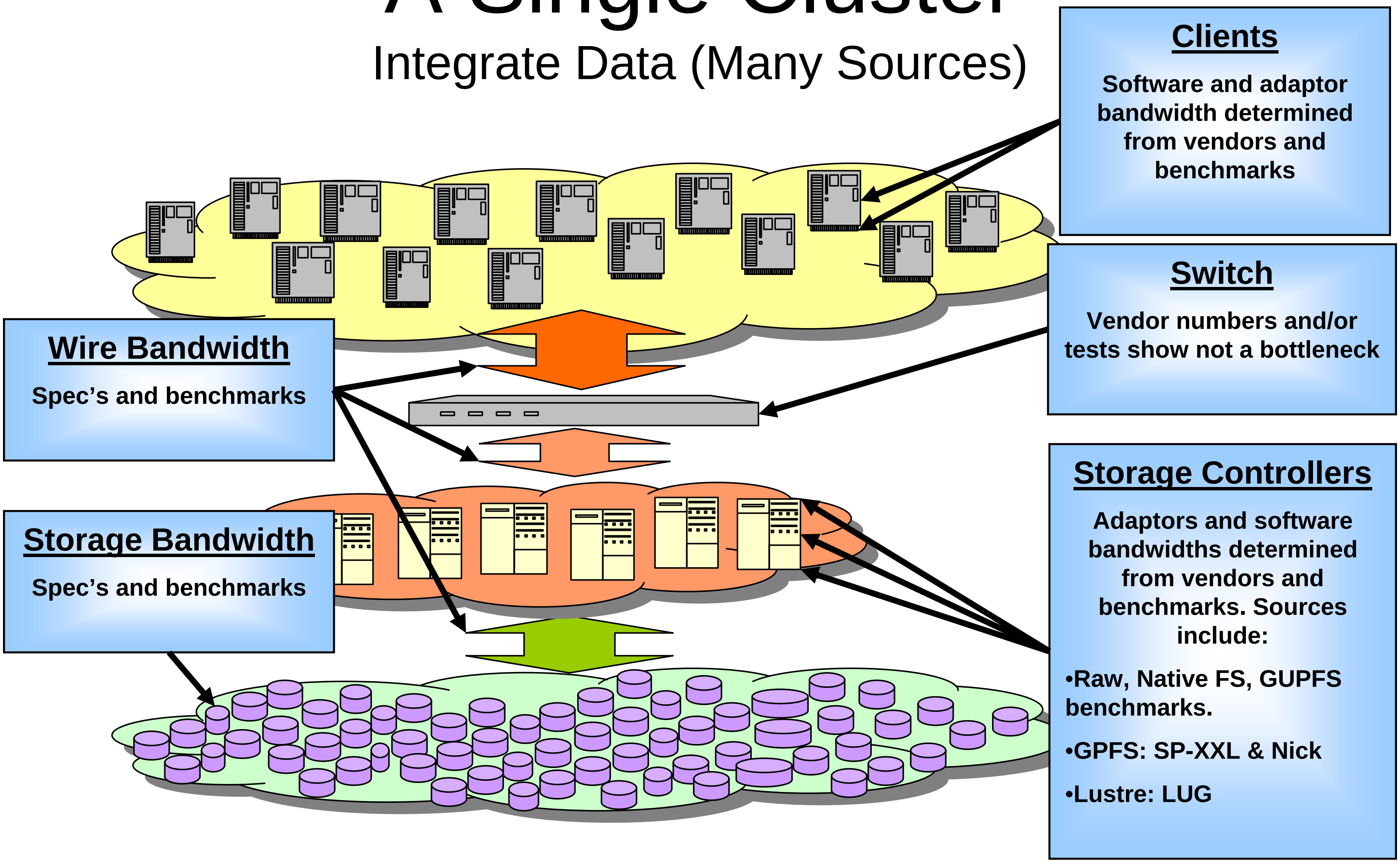
- Goals
  - Integrate Information:
    - Device Characteristics.
    - Number of Components.
    - Bottlenecks.
  - Communicate decisions.
- Define "Costs":
  - Initially defined in terms of <u>time to stream data</u> (like an HPC checkpoint operation).
    - Easy to calculate.
  - Monetary cost can be calculated by adding columns for component pricing.
    - May require vendor interactions (be discreet!)
    - Difficult/laborious to get.
- Fit HPSS into our Model.

A Single Cluster



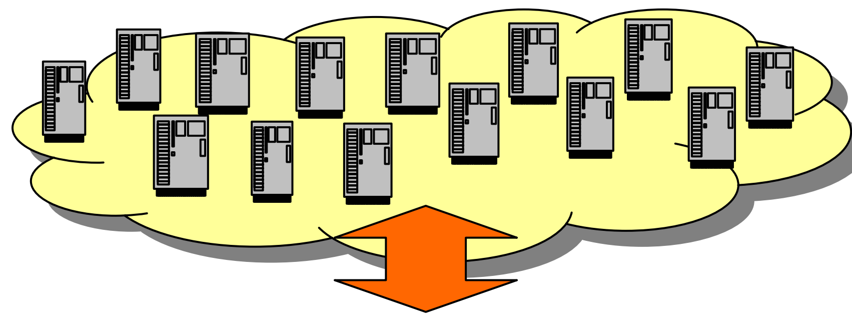
A Single Cluster

Integrate Data (Many Sources)



Single Cluster: Compute Nodes

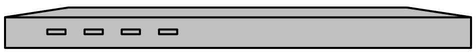
- Walking through an Upgrade Path:
  - 2006: 10GigE is installed.
  - 2007: the number of systems is doubled.
  - 2009: the number is doubled again.



Compute Nodes									
Data Rate Into Switch									
Year	CPUs per node	Amount of Data per CPU to Stream (GB)	Amount of Data per Node to Stream (GB)	Number of Compute Nodes	Aggregate Amount of Data to Stream	Interface Data Rate (Gb/s)	Number of Ports per Interface	Efficiency of Transport	Aggregate Data Rate to Switch (Gb/s)
2004	2	2	**4**	250	**1000**	1	1	70%	**5600**
2005	2	2	**4**	250	**1000**	1	1	70%	**5600**
2006	2	2	**4**	250	**1000**	10	1	70%	**56000**
2007	2	2	**4**	500	**2000**	10	1	70%	**112000**
2008	2	2	**4**	500	**2000**	10	1	70%	**112000**
2009	2	2	**4**	1000	**4000**	10	1	70%	**224000**

Single Cluster: Simple Switch

- More complex switch topologies can be represented.
  - Number connections and throughput can be determined from the spreadsheet.

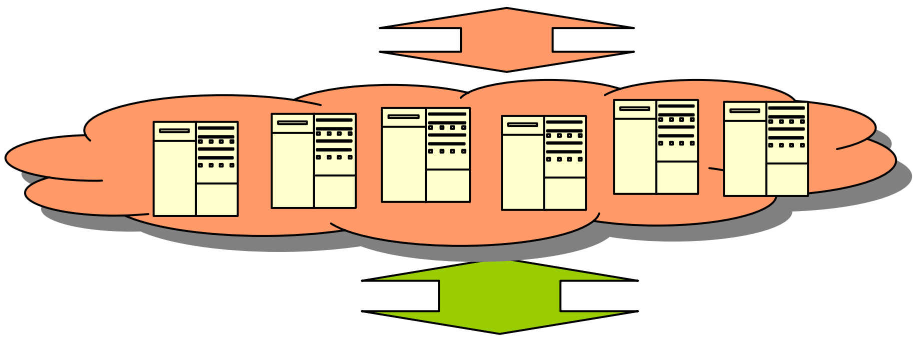


Switch			
Data Rate Through the Switch			
Year	**Switch Fabric Data Rate (Gb/s)**		**Effective Data Rate of Switch (Gb/s)**
2004	10000		**10000**
2005	10000		**10000**
2006	10000		**10000**
2007	10000		**10000**
2008	10000		**10000**
2009	10000		**10000**

Single Cluster: Storage Controllers

Walking through an Upgrade Path:

- Note: in 2004, both the network and storage interfaces are bottlenecks.
- 2005: The number of storage controllers is doubled.
  - 2006: 10GigE is installed.
  - 2007: Faster storage interfaces installed. *Storage Controller throughput is now the limiting factor.*
- 2007: The number of Storage Controllers is doubled.



Storage Controllers												
	Bandwidth Into Controllers						Bandwidth to Storage					
Year	Number of Storage Controllers	Switch Interface Data Rate (Gb/s)	Number of Connections to Switch per Controller	Switch Efficiency of Transport	Aggregate Data Rate to Controllers (Gb/s)	Internal I/O Throughput (MB/s)	Storage Interface Data Rate (Gb/s)	Storage Interface Number of Ports	Storage Efficiency of Transport	Aggregate Storage Interface Data Rate (Gb/s)	Effective Storage Controller Data Rate (Gb/s)
2004	10	1	2	70%	14	500	1	2	70%	14	14
2005	20	1	2	70%	28	500	1	2	70%	28	28
2006	20	10	1	70%	140	500	2	2	70%	56	56
2007	20	10	1	70%	140	500	10	1	70%	140	80
2008	40	10	1	70%	280	500	10	1	70%	280	160
2009	40	10	1	70%	280	500	10	1	70%	280	160

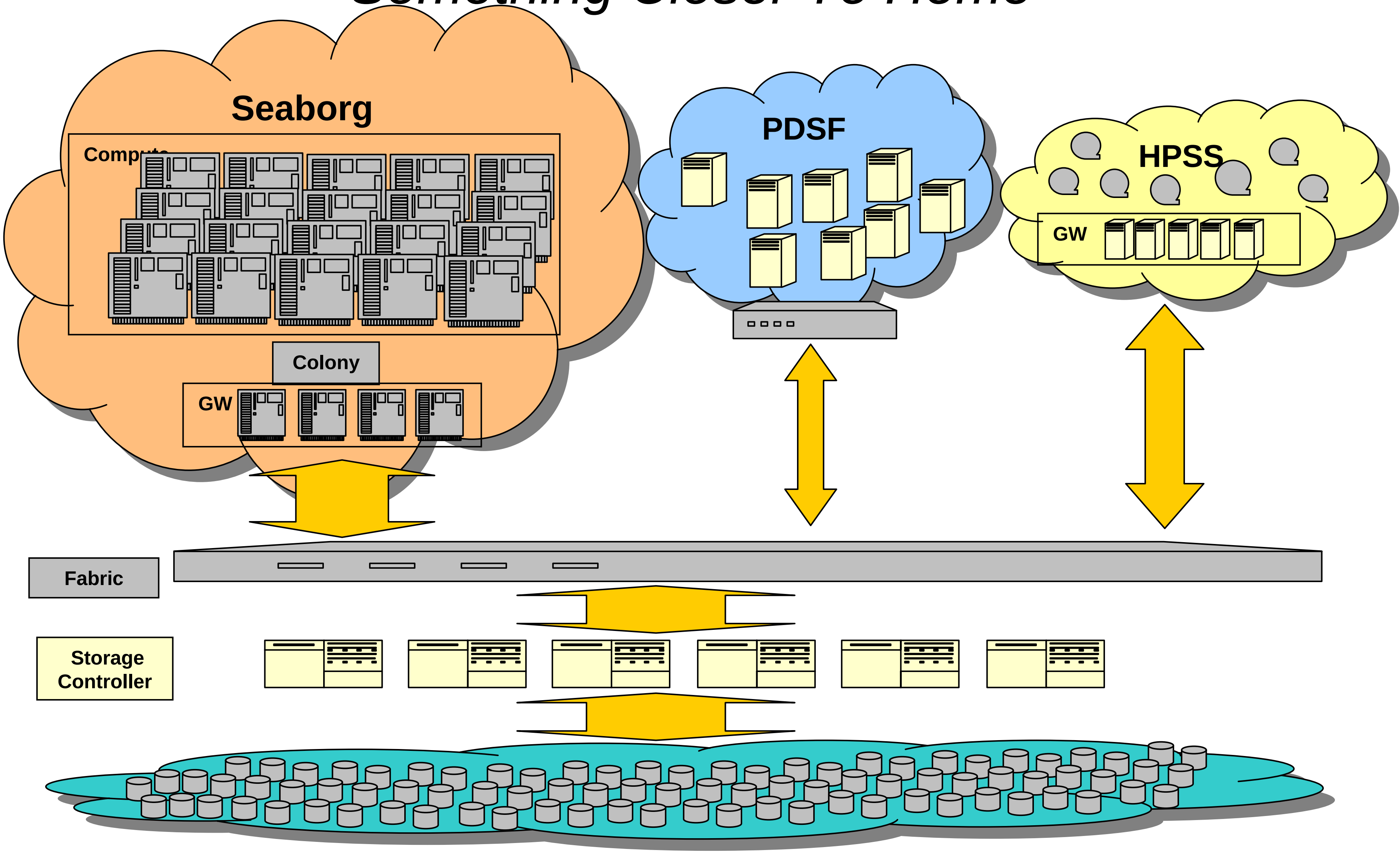
Single Cluster: Center Summary

- In this scenario, the storage controllers are always the bottleneck.
- “Cost” is determined by the time to perform the Streaming IO.

	Centerwide Bandwidth Overview				Time To Perform Streaming I/O			
Year	Compute Throughput (Gb/s)	Fabric	Storage Controller (Gb/s)	Compute to Storage Bandwidth	Amount of Data (GB)	Maximum Time (seconds)	Time (seconds)	Time (hours)
2004	5600	10000	14	14	1000	1800	71.43	0.02
2005	5600	10000	28	28	1000	1800	35.71	0.01
2006	56000	10000	56	56	1000	1800	17.86	0.00
2007	112000	10000	80	80	1000	1800	12.50	0.00
2008	112000	10000	160	160	1000	1800	6.25	0.00
2009	224000	10000	160	160	1000	1800	6.25	0.00

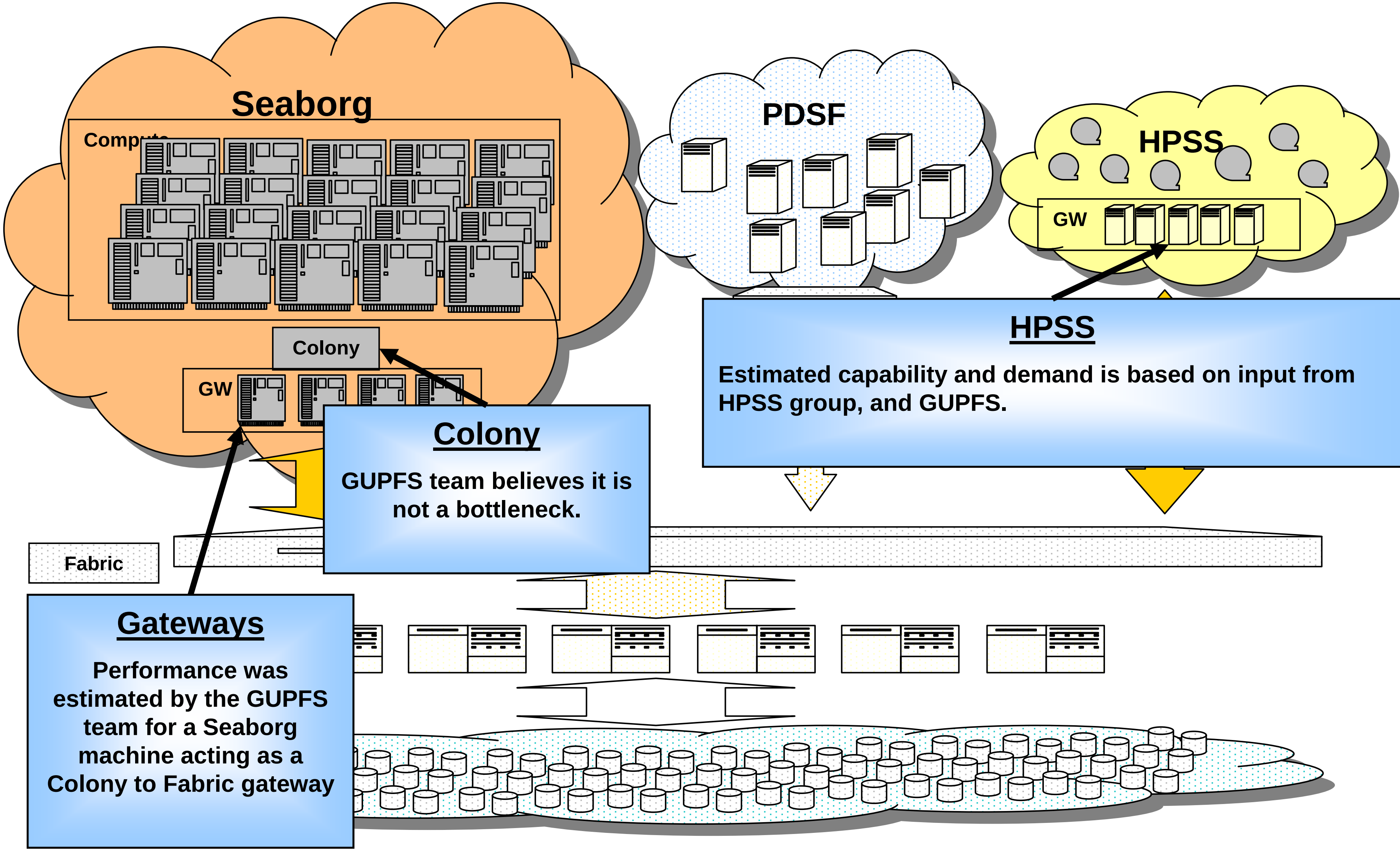
Current Example

Something Closer To Home



Current Example

Integrate Data (Many Additional Sources)



Current Example

Look At Spreadsheet

HPSS Assumptions: Future Capability

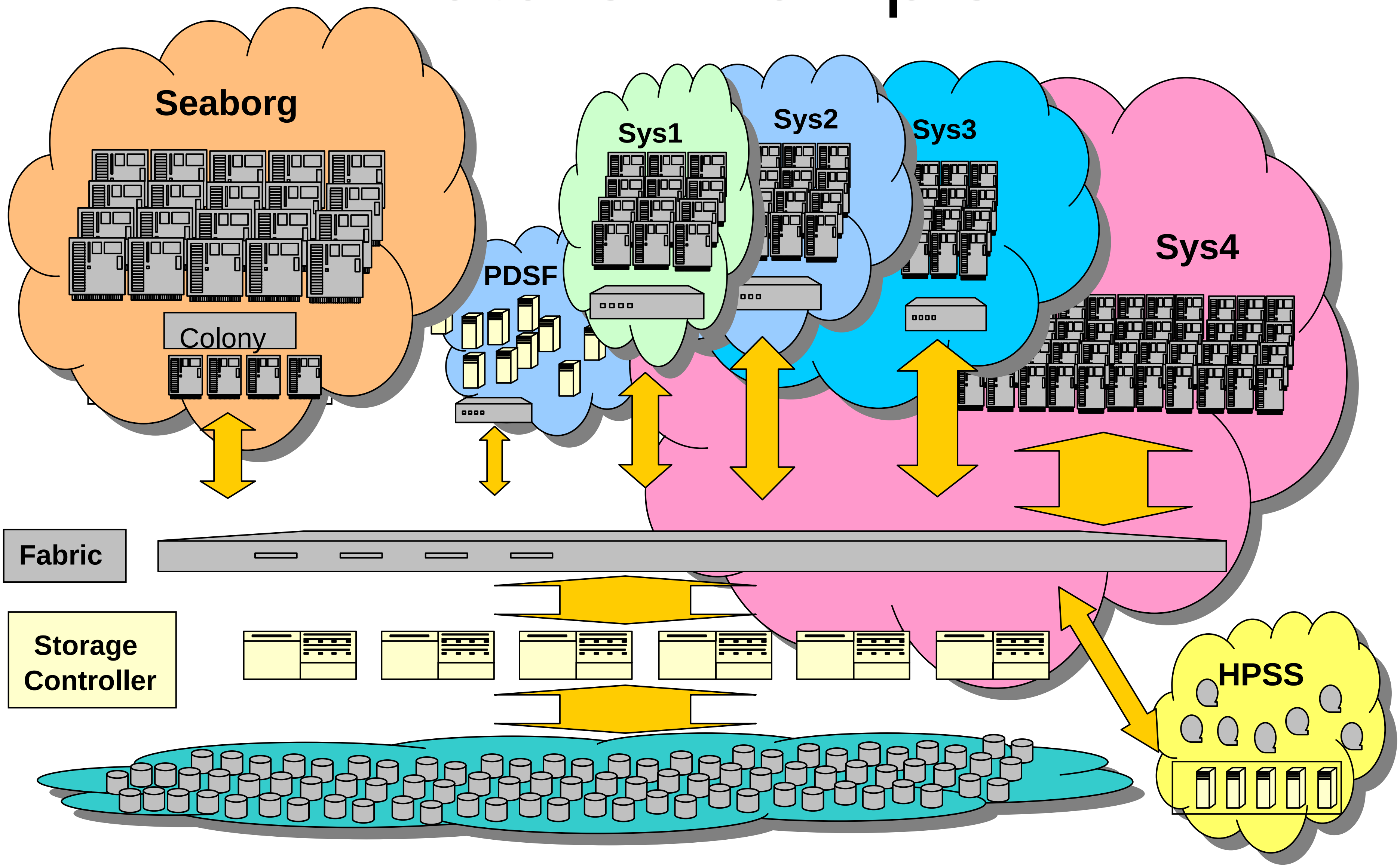
- We note:
  - HPSS performance is limited by tape seek and load times (peak drive performance is quite different from daily observed performance)
  - User activity defines the amount of data to move to/from HPSS.
- Based on discussions with Nancy’s group and within GUPFS, we assume HPSS capability scales as follows:
  - The current HPSS configuration can handle roughly a 3x increase in load or, in other words, a sustained throughput of 10 TB/day.
  - HPSS performance scales according to tape drive performance. For example:
    - Doubling the number of tape drives will permit HPSS to handle twice it’s current and maximum daily throughput.
    - Switching to drives that are twice a fast doubles the amount of data HPSS can move per day.
- We use the latest upgrade roadmap to define future HPSS capability.

HPSS Assumptions: Future Demand

– **We assume future demand on HPSS will scale according to a percentage increase over the current Seaborg usage. For example:**
  - **Adding a cluster which streams the same amount of data to storage as Seaborg currently does will double the amount of data HPSS needs to archive and retrieve.**
  - **Adding four systems with the same IO requirements as Seaborg quadruples the amount of data HPSS must handle.**

– **Assume that switching to an automatic HSM system will not change the average HPSS activity.**
  - **This assumption is actively being discussed.**

Future Example



Future Example

Look At Spreadsheet

Cost Analysis

- Given the component costs, it is possible to price a scenario because:
  - The model uses the number of components (cost*number).
  - Key characteristics can be determined for more detailed specification. For example:
    - Switches can be priced as the model contains numbers of connections at a given speed and total bandwidth through the switch topology.

Summarize/Future Directions

- Provided a bandwidth model for “what-if” GUPFS scenarios
  - Requested by Bill Kramer and James Craw
- Illustrated how we can meet our goals of Integrating and Communicating GUPFS data.
- Made the model available since February to the GUPFS team.
- Intend to make available to facilitate conversations as part of the analysis and procurement process.

Questions?